DOCUMENT RESUME

ED 430 005 TM 029 718

AUTHOR	Hearne, Jill; Klockars, Alan
TITLE	Fairness Issues in Student Exit Performance Assessment.
PUB DATE	1999-04-00
NOTE	18p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Academic Standards; *Bias; *Elementary School Students; Intermediate Grades; *Performance Based Assessment; Reading Achievement; Student Evaluation; *Student Promotion; Urban Schools
IDENTIFIERS	*Exit Examinations; *Fairness; Iowa Tests of Basic Skills

ABSTRACT

A standards-based exit policy was implemented in an urban district in the northwestern United States. This paper considers the second year of implementation with a group of 2,581 students in the fifth grade. Of these, 104 were identified as not having the skills to exit fifth grade. Reading achievement as measured by two tests, the Iowa Tests of Basic Skills and the Washington Assessment of Student Learning, was cross tabulated with teacher judgments. There was over 90% correspondence between the two tests, both of which were published by the same company, and there was a high relationship between teacher judgments and student achievement on external measures. Teachers had high validity in making judgments about student learning relative to their performance on external measures, with good agreement for 90% of students. Some reasons for misjudgments about student achievement are discussed, and the need for additional professional development to ensure equity standards-based exit requirements is emphasized. (Contains 3 tables and 28 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 430 005

Fairness Issues
In
Student Exit Performance Assessment

Jill Hearne
Seattle Public Schools
815 4th Ave. N. Seattle WA 98109
(206) 298-7235
jhearne@cks.ssd.k12.wa.us

Alan Klockars
University of Washington
Miller Hall DQ-12
University of Washington
Seattle, WA 98195
(206) 543-1139

Presented at the American Educational Research Association
Annual Meeting
Montreal, Canada
April, 1999

TM029718

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY
Jill Hearne
TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

BEST COPY AVAILABLE

Fairness in Student Exit
Performance Assessment

Jill Hearne, Ph.D.
Seattle Public Schools
Alan Klockars, Ph.D.
University of Washington

Objectives

Issues of standard setting lie at the intersection of statistical reliability and validity and politics and policy making. Standard setting is no an absolute procedure, but one that has been influenced by history (Madaus,, 1992 and politics (Linn, 1994). What standards (or standards of performance) define the necessary set of skills to be successful at the next level of schooling and/or successful as self-determined citizen in a democracy? What is the similarity (or difference) between these standards and professional and public perceptions of competence? And finally, what is the relationship of teacher judgment of standards attainment and more conventional norm-referenced measures?

These are the questions and issues which surround the design of a high stakes exit policy as one district constructs a model of identification and assistance for all students in attaining standards of proficiency in reading, mathematics, and language arts. This study describes the second phase of an examination of the following questions:

1. What is the similarity of classification of student competency on two differently constructed large scale assessments?

2. What is the relationship between teacher judgments based on classroom evidence and district and state evidence of student achievement?

3. If a disagreement between teacher judgments based on classroom evidence and district/state evidence is called "error", do error rates reflect ethnic or gender disproportionalities?

Perspective

After several turbulent years engaged in issues of decentralization and restructuring, a large urban district is engaged in construction of a centralized accountability policy which depends on school-based implementation. This top-down, bottom-up approach to policy implementation is suggested by Elmore and Associates (1991) as necessary in systems with little centralized power and control over constituents.

The advent of a strong superintendent and a cohesive board made it possible to put into place a policy which holds individual schools responsible for individual student achievement with standards of performance being demonstrated through classroom evidence and supplemented by state and district level assessments.

Inherent in any policy construction surrounding higher standards are the issues of reliability and constancy of teacher judgment. Are teacher judgments about student proficiency fair and equitable to all students or are they tainted by "limiting beliefs about differential ability to learn and self-defeating teaching methods that follow from such beliefs?" (Weinstein, 1996). If there is a differentiated standard that shifts in response to perceived student potential (Weinstein, 1996) then grade advancement decisions that rest on classroom based assessments might increase the likelihood that marginalized groups are overrepresented in the population not advanced. Student achievement can be affected by teacher beliefs when teachers reduce the amount and level of schoolwork given students when this reduction is not necessary (Goldenberg and Gallenmore (1991).

In the 1997-1998 school year these researchers followed the first year of implementing this standards-based exit policy (Hearne and Ramey, 1998). This policy was implemented in this Northwest urban district of approximately 50,000 students in the Spring of 1998. There were 534 students in 4th grade who had been identified as not on track to exit 5th grade. Summer school intervention and school level assistance such as Saturday School and tutoring reduced the number of students identified as not on track to exit. The paper presented at AERA in April, 1998 described the investigation of the application of the policy at various schools and its impact on the students identified as not having the skills to exit.

This paper describes the second year of implementation with a group of 2581 students now in fifth grade. Of these, 104 students were now identified as not having the skills to exit fifth grade.. Fourth grade test scores on two external measures were used to determine whether or not the students

were correctly identified as being on track to exit the grade. Of the 2581 students, 1294 were majority and 1285 were minority, with a balance between males and females.

Among the 104 students not on track to exit, 20 students were majority and 84 students were minority. More males than females were represented in the "not on track" group: within the majority group, 9 females to 11 males; and in the minority group, 30 females to 54 males.

Issues of disproportionately low achievement among minority groups, particularly African American males are of central concern in this district and each school has identified students as either: (a) yes, meets grade level standards or (b) no, does not meet grade level standards. Last year, differences were found in teacher judgement of students meeting standards at the student level, classroom level and school level. Errors were of two types. Type O (for over-expectation) errors occurred when teachers judged students as having skills when the test scores placed him below standard. Type U (for under-expectation) errors occurred when teachers said students did not have skills when their test scores placed them above the minimum level for meeting standards. This year's analysis follows the same questions with the exception of school level analysis which was not possible due to factors external to this study.

Method and Data Sources

Design

The study will use SPSS CROSSTABS to address the above-listed questions.

Grades. The grade under study is 4th grade.

Subject: Reading achievement as measured by the two tests will be crosstabulated with teacher judgement. Reading was selected due to its influence in teachers' making promotion/retention decisions.

Variable 1 is a categorical variable, the end-of-year teacher exit readiness judgments. The two categories of teacher judgments of exit readiness are Yes, meets standards and is projected to pass the exit grade (fifth) and No, does not meet standards and is not projected to pass the exit grade (fifth). Teachers' exit readiness judgments for each student are stored on district computer files.

Value labels are test scores, which are stored by student, on district computer files. There were two tests used. The Iowa Test of Basic Skills is a standardized norm-referenced test. For this study, a

score of 35 Normal Curve Equivalent (NCE) was considered to "meet standard" as this is the score used for exiting students from federal program assistance (Bilingual, Title I, etc).

The second test is the Washington Assessment of Student Learning (WASL). This test is an extended test that combined multiple choice, short answer and extended response items. Scores on the Washington Assessment of Student Learning are reported as scale scores. In creating scale scores the Partial Credit Model (PCM was applied to student-level data. This resulted in an equal interval scale that functions much like a ruler that is marked in inches or centimeters.

A standard setting process called "book-marking" was used to establish levels of performance on this assessment. The scale score range is 150 to 600 and the standard was set at 400 for "proficient". Exemplary work was set at 421 and 375 is the cut point between "beginning" and "developing" work. Students with scores below 375 were judged to be likely to have skills low enough to be of concern for promotion, as "developing" work has evidence of some skill development. The Technical Manual defines level 2 as "....partial accomplishment of the knowledge and skills that are fundamental for meeting the standard at grade 4".

Data:	Grade Four	Spring '98 District	ITBS Reading
		Spring '98 State	WASL Reading
		Teacher Judgement of Proficiency	Scantron Reports

The analysis consists of a series of crosstabulations to examine frequency of error type O and U for the entire population and by gender and ethnicity. Differences in error rate by teacher judgement and by test were examined. Since the teacher judgement variable is dichotomous, it was scored 1 (yes), student has skills to pass or 2 (No, student does not have skills to pass

Classification of cases. Teacher judgement as reported on District scantron sheets is the basis for classifying students. Teacher identification of students as "yes, meets standards and is on track to exit" and "no, does not presently meet standards but is making sufficient progress to exit" were classified as a "1" for pass. Teacher identification of students as "No, is not meeting standards and is not on track to exit" were classified as "2" for fail.

There are two kinds of errors: the first type we will call Type O for over-expectation; classifying a student in the "yes, Pass" category when he has test scores below 375 on the WASL or below 35th NCE on the ITBS; and the second type we will call Type U for under-expectation; classifying a student in the "no, Fail" category when she had test scores of 375 and above on the WASL or 35th NCE on the ITBS. Likewise, there are two kinds of correct classifications. Teacher says "yes, has skill" and the test

corroborates this and conversely the teacher says "no, Fail " and the test evidence supports this judgement.

Question 1: What is the similarity of classification of student competency on two differently constructed large scale assessments?

When the classification rating of 35th NCE on the ITBS was examined in its relationship to the scale score of 375 on the WASL, there was high agreement between the two tests.

The correspondence between the two tests resulted in a 91% agreement that students were either pass or fail. The tests disagreed about 9% of the time, with more students not meeting the ITBS standard (257) than the WASL standard (131).

Table I

WASL/ITBS Correspondence

Using 375 and 35th NCE

	WASL Fail	WASL Pass	
ITBS Fail	79 (3.0%)	178 (7%)	257 (10%)
ITBS Pass	52 (2%)	2272 (88%)	2324 (90%)
	131/ 5.1%	2450/ 94.9%	

Phi=.38881

Question 2. What is the validity of teacher judgements, based on classroom evidence, for decisionmaking in a large urban district's exit profile system?

When teacher judgement about students' ability to meet standards in Reading were analyzed in relationship to students test scores on the WASL, the performance assessment, teachers were very accurate. There was 92.8 %of the students for whom there was agreement between the test and teacher judgement on meeting standards and 1.2% of the students for whom there was agreement between the test and teacher judgement that they did not have the skills.

In combination, the teachers' correct judgements resulted in 94% of the students being correctly identified. For the total population, only 3.7 % students were judged by teachers to have greater skills than these demonstrated on the test while only 2.3 % of the students demonstrated skills on the test that the teachers did not think they had.

Table II.

Comparison of Teacher Judgement and WASL Test Scores

	WASL Fail	WASL Pass	Row total
Teacher fail	37 (1.4%)	67 (2.4%)	104 (3.8%)
Teacher pass	101 (3.7%)	2503(92.4%)	2604 (96.2%)
	138 (5.1%)	2570 (94.9%)	2708 (100%)

On the ITBS, the standardized norm-referenced measure the accuracy rate was less. There was 88% of the students for whom there was agreement between the test and teacher judgement on meeting standards and 2.5% of the students for whom there was agreement between the test and teacher judgement that they did not have the skills.

In combination, the teachers' correct judgements resulted in 91% of the students being correctly identified. For the total population, 7% of the students were judged by teachers to have greater skills than these demonstrated on the test while only 1.6 % of the students demonstrated skills on the test that the teachers did not think they had.

Table II.

Comparison of Teacher Judgement and ITBS Test Scores

	ITBS Fail	ITBS Pass	Row total
Teacher fail	63 (2.5%)	41 (1.6%)	106 (4.1%)
Teacher pass	188 (7.4%)	2226 (88.4%)	2414 (95.9%)
	251 (10%)	2267 (90%)	2518 (100%)

Question #3 If a disagreement between teacher judgments based on classroom evidence and district/state evidence is called "error", do error rates reflect ethnic or gender disproportionalities?

First of all, Many more minority students were judged "no, not meeting standards" by the teacher (84) and many more minority students did not meet the test standard either on both the WASL(106) and the ITBS (201). Judgements by teachers of students' ability to meet standards varied by race and gender. Their teacher error rates varied by test, as well when we examined error rate by group within each decision. Although the WASL agreed best overall with teacher judgement, when teachers did make errors they were more likely to overestimate or underestimate minority student skills

relative to the WASL than they were to misjudge majority skills. For minority students, teachers were more likely to judge students similarly to the ITBS than to the WASL. In judging students reading ability, 5.7% (75/1302) of the minority students who were passed by the teacher failed the WASL

For minority students, the two types of error occurred slightly more relative to the WASL than to the ITBS but teacher judgement also varied more for the WASL than for the ITBS for both majority and minority groups. On the WASL, Type O and Type U error for minority students was 5.7% (75/1302) and 63% (52/83) respectively. Of these Type O errors, , 33 were female and 42 were male. Of the students whom the teacher rated as "not meeting standards, Type U, , 22 were female and 30 were male.

Of the 2% of the majority students who were passed by the teacher but failed the test, 16 were female and 10 were male. The majority students' total failure rate as judged by the teacher was significantly less, only 21 students. Of these 21, however, 15 of them actually met the high standard of the WASL. giving the teacher a comparable error rate of 71%. 6 students were female and 8 were male.

On the ITBS, Type O and Type U error for minority students was 12.5% (151/1201) and 40% (34/84) respectively, while for majority students, the error rates were 3% for Type O (37/1202) and 35% for Type U (7/20). Missing cases account for the change in denominator between the two tests. Thus, the teachers were more likely to identify students as having skills when they did not than they were to identify skills in students they have labeled as not meeting standards.

Table III

Comparison of Minority/ Majority Frequency of Failure and Teacher Judgement

	WASL Below 375	ITBS-Less than 35thNCE	Total Teacher Judged 'No, Not	Type O error* WASL	Type O error* ITBS	Type U error** WASL	Type U error** ITBS
Minority n=1385	106	201	83	5.7%	12.5%	63%	40%
Majority n=1311	32	50	20	2%	3%	71%	35%

4. *reported as % of total not meeting test standard

5. **reported as % of total judged as "No, not meeting..." by teacher

For minority students, the two types of error occurred slightly more for males as for females. On the WASL, for those minority male students who were judged to have skills but did not, 42 out of 626

(Type O= 6.7%) did not meet the test standards, while among those indicated as "no, not meeting standards by the teacher, 30 out of 51 did have skills (Type U= 59%). For minority females the error rates were 5% for Type O and 69% for type U. For majority males, the error rates were: Type O, 1.5%; and Type U, 75%and for majority females, the error rates were Type O=2.5% and Type U=66%. On the ITBS, Type O and Type U error for minority male students was 12% and 40% respectively, while for majority male students, the error rates were 3% for Type O and 45% for Type U. On the ITBS, Type O and Type U error for minority female students was 13% and 40% respectively, while for majority female students, the error rates were 3% for Type O and 22% for Type U .

Table III

Error rate by Gender and Ethnicity on the WASL and the ITBS

	WASL		**ITBS**	
Total Minority error	9%		14.3%	
Total Majority error	3%		3.6%	
Minority male	Type0=6.7%	Type U=59%	TypeO=12%	Type U=40%
Minority female	Type 0=5%	Type U=69%	Type O=13%	Type U=40%
	WASL		**ITBS**	
Majority male	Type O=1.5%	Type U=75%	Type 0=3%	Type U=45%
Majority female	Type O=2.5%	Type U=66%	Type O=3%	Type U=22%

Teachers were more likely to rate positively majority females who did not have test evidence and more likely to rate majority males positively than minority males. Teachers were more likely to over expect than underexpect, but when they did under-expect, as evidenced by their stating students did not have skills when their test scores indicated they were above the minimum levels, it was the ITBS that was more closely aligned to their judgement.

Discussion:

Question #1 seeks the similarity of classification of student competency on two differently constructed large scale assessments. There was over 90% correspondence between the two external measures. The two tests were published by the same company, Riverside Publishing, and have been marketed with the rationale that the ITBS represents the "Basics" while the WASL represents the

"Basics Plus" (Bergeson, 1996). While the ITBS focuses on knowledge, fact and recall questions, the WASL requires students to exhibit writing skills, the ability to synthesize text and to problem solve. Other studies have found a strong correlation between norm referenced tests and the WASL.(Hearne and Ramey, 1998), (Ensign and MacQuarrie, 1998). It would seem that the two tests do tap into similar skills, but at different depths.

Question #2 examined the relationship between teacher judgements based on classroom evidence and district and state evidence of student achievement. Our study found a high relationship between teacher judgement and student achievement on external measures. This finding supports the work discussed in a 1989 meta-analysis of teacher-based judgements of achievement. In an examination of 16 studies investigating the match between teacher based assessments of student achievement levels and objective levels of student achievement, the data revealed high levels of validity for the teacher judgement measures, (Hoge and Coladarci, 1989)

When teachers did make errors, they were more likely to judge that the student did not have skills when they did have them than they were likely to believe the student had skill when they did not. Teachers' type U error rate on the WASL was 64% of those student labeled as "no, nots', as opposed to only 3 % of those students labeled as having skills that didn't. On the ITBS, Teachers' Type U error was 39% of those whom they called "no, nots" versus 8% of those students whom they labeled as ahving skills that didn't. It is possible that teachers saw daily work for these latter students that demonstrated skill. There may well be classroom-based evidence that these students do meet standards particularly for students who were close to either the 35th NCE criterion or the 375 Criterion.

But for some reason there remains a group of students for whom teachers can't see skills in the classroom. While Hoge and Coledurci (1989) concluded that the teacher achievement judgements are generally veridical, they did find that teachers were influenced by their perceptions of student academic ability. They were more likely to correctly identify the performance of their highest quartile students than their lowest quartile students.

Particularly comparing error rates on the ITBS to the WASL, teachers seem to be more accurate at judging basic skills that they are at judging students' higher level responses. This is counter to the historical issues of test bias and disproportionality gap, which maintain that multiple choice tests are less fair to students than judging actual student work (Singham,1998). The classroom-based evidence that these students do meet standards may match more closely the ITBS than the WASL. In these

classrooms daily work may not resemble the kinds of application and analysis required on the WASL but may be more like the multiple choice answers required on the ITBS test.

In further examination of the error rates, Question #3 asks if error rates reflect ethnic or gender disproportionalities.

When we look within the overall error rates we see that there are differences by gender and ethnicity in error rates. Since more minority students did not meet the standard on the external measures and more of then were judged as "no, not meeting standard" by their teacher, there was a larger group of minority students with whom teachers could make errors. Given the issues of disproportionality and test bias, this achievement gap can be expected, according to some (Herrnstein and Murray, 1994). Perhaps the teachers' tendency to misjudge their lower students influenced their decision making process. Also, their error rate is less relative to the ITBS than to the WASL. This may be due to the tendency to basic skills noted by Koretz, Linn, Dunbar and Shepard(1991). They found that some teachers of minority students tended to focus their mathematics and reading curriculum on content specific to the mandated test, thereby limiting the range of instruction made available to minority students to a purely functional level. This would account for why the Type U errors are greater relative to the WASL for this subgroup

In regard to Type O error on both tests, teachers were three to four times more likely to make over-expectations in judgement relative to minority achievement than to majority student achievements. Gay(1990) posits that even when teachers try conscientiously to control their biases, they may still discriminate between culturally different students. Teachers were twice as likely to pass on a minority female student than a majority female who did not have skill relative to the WASL and four times as likely to pass on a minority female than a majority female relative to the ITBS. A similar pattern emerges for minority males, with teachers being four times more likely to pass them along than their majority peers, using either test as the criterion. Early research on teacher student interaction highlighted the phenomenon that in regard to student achievement it is not just the existence of an expectation that causes self-fulfillment, it is the behavior that the expectation produces. Because teachers expect less, students achieve less. (Brophy and Good, 1994)

This indication of low expectations has been well documented in the literature and creates a self-fulfilling prophecy. Bronfenbrenner (1988) refers to it as "Mathew effects", after Mathew 25:29. "those who have will get more until they grow rich while those who have not, will lose even the little they have." Low expectations limit the opportunities for appropriate instructional intervention if a student is

advanced to a higher grade in school without an academic support structure. Those perceived as good readers, writers, and thinkers are provided both increased opportunities to read, write and think critically. Those designated as poor readers or disabled learners are given few opportunities to read, write, or think critically because it is assumed that they are not ready to do what the "able learners" are doing. (Bartoli, 1995)

The more troubling of the two types of error from an ethical point of view is the Type U error rate. Since the Type U error refers to the teacher judgement of students not meeting standards when state and district measures indicate they do have the skills, these students may not be recognized for what they can do. Type U error varied from a low of 22% for majority females to a high of 45% for majority males, relative to the ITBS. Given the small numbers, (2/9) and (7/20) respectively, it is likely that classroom behavior issues were accounting for teacher perception of achievement here. The clustering of Type U error of underexpectation relative to the WASL from 59% to 75% for the subgroups may indicate teachers are looking at something besides skills as exhibited in student work. Perhaps they just don't know these students, don't know what they are doing and thus what they are capable of doing.

The similarity in Type U teacher judgement error for minority and majority students who do have evidence of meeting standard in Reading may reflect a lack of clarity on the teacher's part regarding what constitutes "evidence' of competency. It is difficult to imagine, given the performance level required to score in the "developing" range that the students would not be doing the same kind of work in class, unless, in fact in these classrooms daily work does not resemble the kinds of application and analysis required on the test.

The highest error rates in all categories occurred for minority females. This may also reflect lower expectations on the part of the teacher for females or less ability to judge their work as a 'stand alone' and not as an adjunct of behavior. The WASL reading assessment includes some multiple choice items but includes two extensive constructed response tasks that require analysis and synthesis of information. It is highly unlikely that students could overscore on this assessment. 22 out of the 32 student not passed by the teacher did exhibit these skills. If girls can read, but their teachers indicate they cannot read, it points to the phenomenon studied by Sadker (1994), Oakes (1990) and others. Harvey (1986) and Sadker and Sadker (1986) reported that minority females receive the least attention in the classroom and most teachers are not aware of their own inequitable interactions with females. Receiving less attention, females are less likely than males to have opportunities to respond to open-

ended questions and exhibit higher order thinking skills. Harvey and the Sadkers found that brief focused teacher training can reduce or eliminate these inequities—which underscores the unintentionally of this inequity.. This training is essential because if teachers do not expect that students can take part in a higher level discussion, those students are not even given a chance to participate (Stallings and McCarthy, 1990).

The prevailing wisdom is that performance assessment will more closely match student skill demonstrations with that which is done in the classroom, but it may be that in classrooms where students are predominantly minority, the curriculum resembles more closely the knowledge, fact and recall level of the ITBS rather than the application, analysis, and synthesis level of the WASL. If teachers are less able to recognized skill development as evidenced in performance assessment, then they are also less likely to prescribe appropriate strategies for learning. Indeed, Black writes that "...teachers are often able to predict pupils results on external tests because their own tests imitate them, but at the same time teachers know too little about their pupils' learning needs. (1998). Thus we see a smaller error rate of both types for the ITBS instead of the WASL, even though initially the ITBS has more students not meeting standards.

This can be further understood in the light of the historical issues of test bias and disproportionality gap. Minority students may be under-performing in class as a result of many issues ranging from fear of being rejected if outperforming peer 'expected' levels or their perception of the relationship between effort and reward (Singham, 1998). If so, the teacher may not be "seeing" the skills embedded in student daily work.

Conclusions and Recommendations

At the conclusion of the second year of this large urban district's implementation of an exit profile system, teachers have high validity in making judgements about student learning relative to their performance on external measure. Equity issues in standard setting lie not so much in the standards themselves but in the implementation of those standards. Their clarity provides the opportunity for equitable educational advancement regardless of race and gender only if all decision-makers can accurately judge those who reach standards and assist those who don't.

The training that has occurred for staff during the 1997-99 time period has focussed on developing teachers' understanding of quality classroom based evidence and on learning to judge student work

relative to standards. This work is still in its infancy and fewer than 10% of the teaching corps has been exposed to this kind of staff development to date.

Sarason writes "it is far easier to deal with villains than with well-intentioned educators imprisoned in tradition and by orientations that render self-scrutiny extraordinarily difficult." He acknowledges that teaching s a "taxing, frustrating, satisfying , mind-bending and mind-altering role for those who have not fallen prey to apathy and routine'. (In Bartoli 1995, ix.). In this district this is particularly true at the present time due to the death of a visionary superintendent and subsequent complete change in leadership at the central office level. Due to leadership instability and an over reliance in the past with external norm-referenced tests, it is likely that teachers in this district have not had reason to develop skills or confidence in their ability to judge student work.

Assessment measures must be developed that can illuminate the special talents of students of different ethnic cultural and linguistic backgrounds (Lomax, et al.1996). Teachers must use these measures to increase expectations for underserved groups. In addition, thoroughly communicated district-wide grade level standards are being developed and should be implemented in the next school year. This should help clarify expectations for students and teachers alike.

If as Gay (1990, p.227) maintains, inequities are transmitted through irrelevant test content, testing structures and styles, teacher attitudes, instructional quality and program concentrations the all of these transmission points must be addressed. It is this cluster of transmissions that create an "ecology of inequity". The reversal of this ecology can only come about through extensive staff development, not through "narrow-minded accountability measures that encourage blame placing and denial of individual responsibility.(Bartoli, 1995, p.139)

We must expand teachers' capacity to see skills embedded in student work and collect classroom based evidence that reflects district and stade standards. We must also encourage schools to examine multiple forms of data and multiple representations of student work. No one piece of work or one test score can be a determinant of student progression in grade (Carter, 1952).

We must also expand teachers strategies for providing opportunities to learn to all students regardless of race, class and gender. Understanding that the application of bias is unconscious we should provide staff development alternative teaching strategies such as cooperative learning, role playing, tutoring, team learning, demonstrating, coaching, problem solving, and non-directive teaching. Staff development should also be targeted toward helping teachers understand better how culturally different students go about the process of learning and demonstrating what they know. Teaching

teachers to design, evaluate, and use alternative evaluation techniques should also increase accuracy in judging student work (Gay, 1990). Since teachers may also be reacting to student behavior, instead of student skill, staff development should also focus on management skills and motivation as well.

Continuing evaluation of the exit profile system is necessary to ensure equitable implementation and reduce disproportionality. Ultimately any policy construction regarding large scale application of standards in a high stakes system requires a definition of standard which includes "Opportunity of success"(Phillips, 1996). This requires that standards be a guarantee of standardized conditions that ensure that no students receive an unfair advantage or penalty rather than a guarantee of equal outcomes. We must focus on a renew of instructional strategies and assessment, examination of personal assumption and biases and reconnecting to families and communities..

High standards can provide that focus, and for the more than ninety per cent of students for whom the teacher matches perception and external measure, these standards work. But for the students who are misjudged, either higher or lower than their capacity, we need to advocate. It is up to us to make sure high standards are for everyone.

References

Bartoli, J. (1995) Unequal Opportunity. New York: Teachers College Press

Bergeson, T. "Washington's Accountability System" presentation for the Commission on Student Learning, Kent, WA, 1998,

Black, P "Inside the Black Box- Raising Standards through Classroom Assessment" *Phi Delta Kappan*. October, 1988 (80:2) pp.139-148

Bronfenbrenner, U. (1988). Foreward. In A. Pence (Ed.) Ecological Research with Children and Families New York: Teachers College Press.

Brophy, J. and Good, T. (1994). Looking in Classrooms. New York: Harper Collins

Carter, R. (1952) "How invalid are marks assigned by teachers?" *Journal of Educational Psychology* (43: 218-228)

Elmore, R. and Associates (1991). Restructuring Schools: The Next Generation of School Reform. San Francisco: Jossey-Bass

Ensign, G. and MacQuarrie, D. (1998, December) *1998 Results from the Washington Assessment of Student Learning(WASL)*. Paper presented at the Annual Meeting of the Washington Education Research Association. 1998

Gay, Geneva. (1990). "Teacher Preparation for Equity". In Baptiste, H.P., Waxman, H.C., Wolkende, F. and Anderson, J. (Eds.), Leadership, Equity, School Effectiveness. Newbury Park, CA: Sage

Goldenberg, C. and Gallemore, R. (1991) "Local Knowledge, Research Knowledge and Educational Change: A case study of early Spanish reading Improvement." Educational Researcher 20(8), 2-14.

Harvey, G. (1986). "Finding reality among the Myths: Why what you thought about sex equity in education isn't so". *Phi Delta Kappan* 67(7) 509-512.

Hearne, J. and Ramey, M. (1998, April) *Addressing Validity Issues in Student Exit Performance Assessment* paper presented at the American Educational Research Association Annual Meeting

Herrnstein, R. and Murray, C. *The Bell Curve* (New York: Free Press, 1994

Hoge, R. and Coladarci, T. "Teacher-based judgements of Academic Achievement: A Review of Literature" *Review of Educational Research*, Fall, 1989 (59:3) pp.297-313

Koretz, D.M., Linn, R.L., Dunbar, S.B., and Shepard, L.A. (1991, April) *The Effects of High Stakes Testing on Achievement: preliminary findings about generalizations across tests*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Linn, R. (1994, October). *The likely impact of performance standards as a function of uses: From rhetoric to sanctions.* Paper presented at the Joint Conference on Standard Setting for Large Scale Assessments, Washington D

Lomax, R., West, M.M., Harmon, M.C., Viator, K.A. and Madaus, G. (1995) "The Impact of Mandated Standardized testing on Minority Students". *Journal of Negro Education* (64:2).

Madaus, G. (1992*). A National Testing System: Manna from Above?* Chestnut Hill, MA:Boston College, Center for the Study of testing, Evaluation, and Educational Policy.

Oakes, Jeannie. (1990). Multiplying Inequalities: The effects of race, social class, and tracking on opportunities to learn Mathematics and Science. Santa Monica, CA. Rand Corp.

Phillips, S. (1996). "Legal defensibility of Standards: Issues and Policy Perspectives" *Educational Measurement* (15:2)pp.5-13

Sadker, M. and Sadker, D.. (1986). "Sexism in the classroom: from Grade school to Graduate School". *Phi Delta Kappan* (67:7), 512-515.

Sadker, M. (1994). Failing at Fairness: how America's schools cheat girls. New York: C. Scribner's Sons.

Singham, M. "The Canary in the Mine- the Achievement Gap between Black and White Students" *Phi Delta Kappan* Sept, 1998(80:1)pp.9-15

Stallings, J. and McCarthy, J. (1990) "Instruction that Enhances Equity". In Baptiste (Ed.) School Effectiveness

Stiggins, R.J. (1997). Student-centered Classroom Assessment (2nd ed.). Upper Saddle River, NJ: Prentice Hall

Washington Commission on Student Learning (1997) *Primary Classroom-based Assessment Tool Kit*. Olympia, WA.

Washington Commission on Student Learning (1998) *Technical Information: Washington Assessment of Student Learning- Early Grades* , Olympia, WA

Weinstein, R. (1996) "Higher Standards in a tracked system of Schooling". *Educational Researcher* (26:8) pp.16-18